FGX

Purutçuoğlu, Vilda; Wit, Ernst

Published in:
Biostatistics

Publication date:
2007

Biostatistics (2007), **8**, 2, *pp.* 433–437
doi:10.1093/biostatistics/kxl020
Advance Access publication on August 24, 2006

FGX: a frequentist gene expression index for Affymetrix arrays

VILDA PURUTÇUOĞLU, ERNST WIT*

Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK
e.wit@lancaster.ac.uk

SUMMARY

We consider a new frequentist gene expression index for Affymetrix oligonucleotide DNA arrays, using a similar probe intensity model as suggested by Hein *and others* (2005), called the Bayesian gene expression index (BGX). According to this model, the perfect match and mismatch values are assumed to be correlated as a result of sharing a common gene expression signal. Rather than a Bayesian approach, we develop a maximum likelihood algorithm for estimating the underlying common signal. In this way, estimation is explicit and much faster than the BGX implementation. The observed Fisher information matrix, rather than a posterior credibility interval, gives an idea of the accuracy of the estimators. We evaluate our method using benchmark spike-in data sets from Affymetrix and GeneLogic by analyzing the relationship between estimated signal and concentration, i.e. true signal, and compare our results with other commonly used methods.

Keywords: Affymetrix; Fisher information matrix; GeneChip; Gene expression; Maximum likelihood; Probe-level analysis; Spike-in data sets.

1. INTRODUCTION

High-dimensional Affymetrix oligonucleotide DNA arrays are widely used in biomedical research. Each oligonucleotide probe consists of a small string of DNA 25 base pairs specific for a gene or expressed sequence tag and is immobilized on a glass slide or array. To measure the amount of transcribed RNA, the gene targets are labeled with a dye and hybridized to the probes on the array. Each gene is defined by a set of 11 to 20 "probe pairs," coming from different parts of the gene's DNA sequence. There are two components to each probe pair: the "perfect match" (PM) measures the amount of transcribed perfectly matched target complementary to the mRNA of a specific gene, whereas the "mismatch" (MM) is supposed to measure the amount of nonspecific binding of the target by changing the 13th base pair of the probe. In order to denote the estimated gene expression level, the term "gene expression index" is widely used. There are several methods that are often used for estimating gene expression levels: MAS 5.0 (Hubbell *and others*, 2002), MBEI (dChip) (Li and Wong, 2001), RMA (Irizarry *and others*, 2003), GC-RMA (Wu *and others*, 2004), BGX (Hein *and others*, 2005), mgMOS (Liu *and others*, 2005; Milo *and others*, 2003), and multi-mgMOS (Liu *and others*, 2005).

*To whom correspondence should be addressed.

The GC-RMA method, an extension of RMA, is the first method to consider the idea that besides nonspecific hybridization, MM values also contain some information about the true signal S whose contamination is a fraction p. However, in practice the calculations are simplified by assuming that p is zero. The BGX method actually estimates p from the data. For estimation, a Bayesian hierarchical setup in conjunction with an Markov chain Monte Carlo approach is used. The Bayesian gene expression index is computed as the median of the posterior signal distribution, although Bayesian credibility intervals are also available.

In this paper, we consider the application of a maximum likelihood (ML) alternative to estimate the true signals under a PM and MM intensity model that is similar to BGX. In this way, we intend to reduce computational cost considerably and to maintain the efficiency of the estimators. The website http://www.maths.lancs.ac.uk/∼wite/research contains both the R-code for frequentist gene expression (FGX) function as well as other supplementary material.

2. MODEL FORMULATION AND INFERENCE

A model that induces a relationship between PM and MM probes is one where both PM and MM share part of a common signal S as well as a large nonspecific hybridization component H as an offset term. Assuming lognormality for the probe intensities deals largely with the full extent of the variance heterogeneity across the intensity range, $\log \mathrm{PM}_{ij} \sim N(S_i + \mu_{\mathrm{H}}, \sigma^2)$ and $\log \mathrm{MM}_{ij} \sim N(pS_i + \mu_{\mathrm{H}}, \sigma^2)$, where $j = 1, \ldots, m$ is the probe indicator, S_i is the true expression value for gene $i = 1, \ldots, n$, p is the fraction of "specific" hybridization of the MM probe, and μ_{H} is the mean of the nonspecific hybridization random effect. The constant variance term $\sigma^2 = {\sigma_\varepsilon}^2 + \sigma_{\mathrm{H}}^2$ is the sum of a measurement error term ${\sigma_\varepsilon}^2$ and a nonspecific hybridization term σ_{H}^2. Since the variance components cannot be identified separately, we estimate a combined σ^2 term.

As averages of the log-transformed PM and MM probes are sufficient statistics for their associated underlying means and because analysis of Affymetrix data typically takes place on a probe set level, rather than an individual probe level, we consider that the available data typically consist of $\mathrm{PM}_i := \sum_{j=1}^m \log \mathrm{PM}_{ij}/m$ and $\mathrm{MM}_i := \sum_{j=1}^m \log \mathrm{MM}_{ij}/m$, such that $\mathrm{PM}_i \sim N(S_i + \mu_{\mathrm{H}}, \sigma^2/m)$ and $\mathrm{MM}_i \sim N(pS_i + \mu_{\mathrm{H}}, \sigma^2/m)$. The aim is to obtain estimates for the parameters p, S_i, and μ_{H} by using ML. The loglikelihood l conditional on $\mathrm{PM} = (\mathrm{PM}_1, \ldots, \mathrm{PM}_n)$ and $\mathrm{MM} = (\mathrm{MM}_1, \ldots, \mathrm{MM}_n)$ is given as

$$l(\underline{S}, \mu_{\mathrm{H}}, p|\mathrm{PM}, \mathrm{MM}) = n\ln(m) - n\ln(2\pi) - 2n\ln(\sigma) - \frac{m}{2\sigma^2}\sum_{i=1}^{n}[(\mathrm{PM}_i - S_i - \mu_{\mathrm{H}})^2 + (\mathrm{MM}_i - pS_i - \mu_{\mathrm{H}})^2].$$

The maximum likelihood estimators (MLEs) of the unknown parameters are solutions of the partial derivatives of l equated to zero and of the parameters μ_{H} and S are the explicit functions of the intensities and the MLE of p,

$$\hat{\mu}_{\mathrm{H}} = (\overline{\mathrm{PM}}\hat{p} - \overline{\mathrm{MM}})/(\hat{p} - 1) \quad \text{and} \quad \hat{S}_i = (\mathrm{PM}_i + \hat{p}\mathrm{MM}_i - (1 + \hat{p})\hat{\mu}_{\mathrm{H}})/(1 + \hat{p}^2),$$

where $\overline{\mathrm{PM}} = \sum_{i=1}^n \mathrm{PM}_i/n$ and $\overline{\mathrm{MM}} = \sum_{i=1}^n \mathrm{MM}_i/n$. In order to obtain the MLE of p, we substitute $\hat{\mu}_{\mathrm{H}}$ and $\hat{S}_i$ into $\partial l/\partial p = -m/(2\sigma^2)\sum_{i=1}^n[-2S_i(\mathrm{MM}_i - pS_i - \mu_{\mathrm{H}})]$, which gives a fourth-order polynomial equation which can be written as

$$(\hat{p} - 1)(E\hat{p}^2 + F\hat{p} + G)/(1 + \hat{p}^2) = 0,$$

where $E = n\overline{\mathrm{PM}}\,\overline{\mathrm{MM}} - \sum_{i=1}^n \mathrm{PM}_i\mathrm{MM}_i$, $F = \sum_{i=1}^n \mathrm{MM}_i^2 - \sum_{i=1}^n \mathrm{PM}_i^2 - n\overline{\mathrm{MM}}^2 + n\overline{\mathrm{PM}}^2$, and $G = \sum_{i=1}^n \mathrm{PM}_i\mathrm{MM}_i - n\overline{\mathrm{PM}}\,\overline{\mathrm{MM}}$. There are three solutions for this equation. $\mathrm{SS}_{\mathrm{PM,MM}} > 0$, i.e. a positive

correlation between all the PM and MM signals, the maximum of l is found at

$$\tilde{p} = \frac{(\mathrm{SS_{MM}} - \mathrm{SS_{PM}}) + \sqrt{(\mathrm{SS_{PM}} - \mathrm{SS_{MM}})^2 + 4(\mathrm{SS_{PM,MM}})^2}}{2\mathrm{SS_{PM,MM}}},$$

where $\mathrm{SS_{PM}} = \sum_{i=1}^{n}(\mathrm{PM}_i - \overline{\mathrm{PM}})^2$, $\mathrm{SS_{MM}} = \sum_{i=1}^{n}(\mathrm{MM}_i - \overline{\mathrm{MM}})^2$, and $\mathrm{SS_{PM,MM}} = \sum_{i=1}^{n}(\mathrm{PM}_i - \overline{\mathrm{PM}})(\mathrm{MM}_i - \overline{\mathrm{MM}})$. And in that case, the MLE of p is given as $\hat{p} = \max\{0, \min\{\tilde{p}, 1\}\}$. Note that if $\mathrm{SS_{PM,MM}} \leqslant 0$, then there is no evidence in the data that the MM probes contain any information about the underlying signal. In that case, the estimate of p should be set to $\hat{p} = 0$.

We note that for estimation of the variance terms, the probe means, PM_i and MM_i, are not sufficient. The loss of information can be regained by reconstructing the likelihood function in terms of all data after estimating the parameters S_i, p, and μ_H and to calculate an MLE for σ^2 conditional on the estimates $\hat{S}_i$, $\hat{p}$, and $\hat{\mu}_\mathrm{H}$, giving $\hat{\sigma}^2 = \frac{1}{2nm}\sum_{i=1}^{n}\sum_{j=1}^{m}[(\mathrm{PM}^*_{ij} - \hat{S}_i - \hat{\mu}_\mathrm{H})^2 + (\mathrm{MM}^*_{ij} - \hat{p}\hat{S}_i - \hat{\mu}_\mathrm{H})^2]$.

Asymptotically, MLEs are fully efficient, i.e. they are unbiased and have minimum variance bounds. For a finite number of samples, the covariance matrix of the MLEs is given by I^{-1}, where I is the "observed Fisher information matrix"

$$I = \begin{bmatrix}
\frac{2nm}{\sigma^2} & \frac{m\sum_{i=1}^{n} S_i}{\sigma^2} & \frac{m(1+p)}{\sigma^2} & \frac{m(1+p)}{\sigma^2} & \cdots & \frac{m(1+p)}{\sigma^2} \\
\frac{m\sum_{i=1}^{n} S_i}{\sigma^2} & \frac{m\sum_{i=1}^{n} S_i^2}{\sigma^2} & \frac{m(2pS_1+\mu_\mathrm{H}-\mathrm{MM}_1)}{\sigma^2} & \frac{m(2pS_2+\mu_\mathrm{H}-\mathrm{MM}_2)}{\sigma^2} & \cdots & \frac{m(2pS_n+\mu_\mathrm{H}-\mathrm{MM}_n)}{\sigma^2} \\
\frac{m(1+p)}{\sigma^2} & \frac{m(2pS_1+\mu_\mathrm{H}-\mathrm{MM}_1)}{\sigma^2} & \frac{m(1+p^2)}{\sigma^2} & 0 & \cdots & 0 \\
\frac{m(1+p)}{\sigma^2} & \frac{m(2pS_2+\mu_\mathrm{H}-\mathrm{MM}_2)}{\sigma^2} & 0 & \frac{m(1+p^2)}{\sigma^2} & \cdots & 0 \\
\frac{m(1+p)}{\sigma^2} & \frac{m(2pS_3+\mu_\mathrm{H}-\mathrm{MM}_3)}{\sigma^2} & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{m(1+p)}{\sigma^2} & \frac{m(2pS_n+\mu_\mathrm{H}-\mathrm{MM}_n)}{\sigma^2} & 0 & 0 & \cdots & \frac{m(1+p^2)}{\sigma^2}
\end{bmatrix},$$

where the first and second column belong to the μ_H and p terms, respectively, and the remaining columns denote the terms belonging to the signals $(S_1, \ldots, S_n)$. To obtain the inverse of I, we partition the matrix after the second row and second column as given below in which A is the 2×2 submatrix at the top of the left-hand side of I, B is the $2 \times n$ submatrix at the top of the right-hand side of I, accordingly, and C is the $n \times n$ diagonal submatrix at the bottom of the right-hand side of I,

$$I = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}, \quad \text{so that } I^{-1} = \begin{bmatrix} P & Q \\ Q^T & R \end{bmatrix},$$

in which $P = (A - BC^{-1}B^T)^{-1}$, $Q = -(C^{-1}B^T P)^T$, $R = C^{-1} - C^{-1}B^T Q$. Explicit formulas for the variances of $\hat{S}_i$, useful for confidence intervals, can be found from the diagonal of R, i.e.

$$\begin{aligned} V(\hat{S}_i) = \frac{\hat{\sigma}^2}{mC_0(1+\hat{p}^2)^3}\Bigg[& C_0(1+\hat{p}^2)^2 - (1+\hat{p})^2\left(\sum_{k=1}^{n}(2\hat{p}\hat{S}_k + \hat{\mu}_\mathrm{H} - \mathrm{MM}_k)^2 - (1+\hat{p}^2)\sum_{k=1}^{n}\hat{S}_k^2\right) \\ & - 2(1+\hat{p})(2\hat{p}\hat{S}_i + \hat{\mu}_\mathrm{H} - \mathrm{MM}_i)\left((1+\hat{p}^2)\sum_{k=1}^{n}\hat{S}_k - (1+\hat{p})\sum_{k=1}^{n}(2\hat{p}\hat{S}_k + \hat{\mu}_\mathrm{H} - \mathrm{MM}_k)\right) \\ & + (2\hat{p}\hat{S}_i + \hat{\mu}_\mathrm{H} - \mathrm{MM}_i)^2(n\hat{p}^2 - 2n\hat{p} + n)\Bigg], \end{aligned}$$

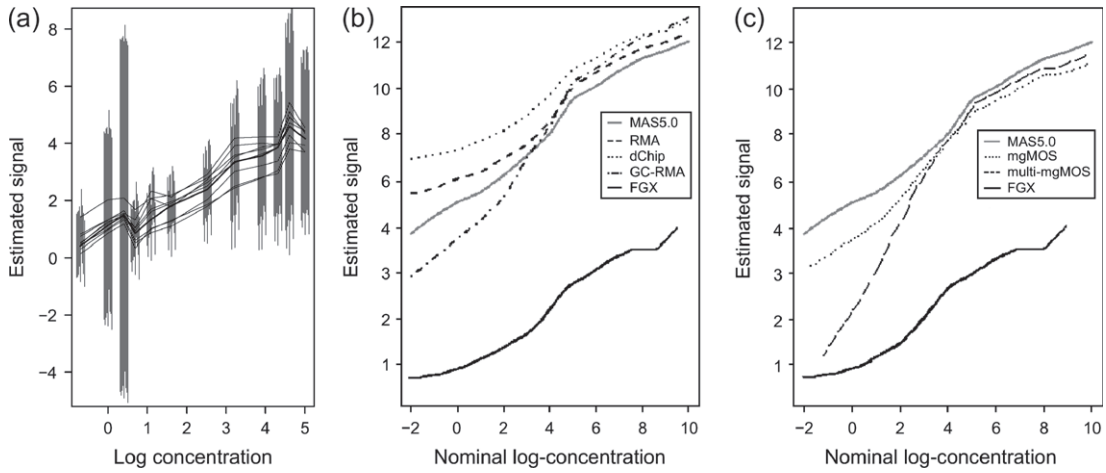

Fig. 1. (a) FGX estimated signal versus true signal and pointwise 95% confidence interval for 10 GeneLogic spike-in genes (excluding array 92 466, spike-in gene DapX-M, and 0.0, 0.75 pM concentrations). (b) Average intensities of 14 Affymetrix spike-in genes versus nominal log-concentrations (excluding 0.0 pM concentration) using MAS 5.0, dCHIP, RMA, GC-RMA, and FGX method, (c) using MAS 5.0, mgMOS, multi-mgMOS, and FGX method.

where

$$C_0 = \left[2n - \frac{n(1+\hat{p})^2}{(1+\hat{p}^2)}\right]\left[\sum_{i=1}^{n}\hat{S}_i^2 - \frac{\sum_{i=1}^{n}(2\hat{p}\hat{S}_i + \hat{\mu}_{\mathrm{H}} - \mathrm{MM}_i)^2}{1+\hat{p}^2}\right] - \left[\sum_{i=1}^{n}\hat{S}_i - \frac{1+\hat{p}}{1+\hat{p}^2}\sum_{i=1}^{n}(2\hat{p}\hat{S}_i + \hat{\mu}_{\mathrm{H}} - \mathrm{MM}_i)\right]^2.$$

3. APPLICATION

We compare FGX and BGX using the GeneLogic spike-in hgu95a data set, which is available from http://www.genelogic.com/newsroom/studies/index.cfm. The GeneLogic spike-in data set consists of 14 arrays with 11 GeneLogic spike-in probe sets and each probe set consists of 20 probes. Apart from gene CreX-3, each of the 10 spike-in genes is hybridized at various concentrations from 0.0 to 150.0 pM. All computations were in R using the affy and hgu95acdf packages for importing and handling the data.

Despite the structural similarities between the FGX and BGX models, FGX slightly outperforms BGX in terms of the so-called "slope detect," in that the slope 0.60 (0.77, if omitting low concentrations) calculated from regressing expression values on the log-concentrations are somewhat closer to 1 than the 0.50 (0.60) achieved by BGX (Hein *and others*, 2005, Figure 6). We note that only a slope of 1 on the log-scale corresponds to a linear relationship on the original scale. From the plots of FGX signals in Figure 1(a), it is seen that the weighted average of estimated FGX intensities of each array have larger confidence intervals than those of BGX at high concentrations (Hein *and others*, 2005, Figure 6).

At low concentrations, on the other hand, both methods have large confidence interval since low measurements are highly affected by the noise and the nonspecific hybridization. Most importantly, due to its explicit formulas the FGX is much faster (1 s in R) than the BGX method (70 min in C++).

In order to compare the performance of FGX to other benchmark methods, we use the Affymetrix spike-in data, available from http://affycomp.biostat.jhsph.edu/. This data set involves in 59 arrays with

14 spike-in probe sets, when excluding anomalous probe sets 33 818 and 546. Each probe set has 16 probes. These spike-in genes are measured at concentrations from 0.0 to 1024.0 pM. In Figure 1(b, c), the average estimated signal across all genes is plotted against nominal log-concentration level. It is interesting to note that FGX is the only method that estimates an effectively zero signal, when the concentrations are negligible. The reason is that by assuming that the MM probe contains some level of the true signal, the average level of nonspecific signal μ_H is identifiable in the FGX model. The R-squared obtained by FGX from regressing the expression values on the log-concentrations is considerably higher (0.95) than those obtained by the other methods (0.80–0.86). However, apart from these two points, which clearly pack out in favor of FGX, there is a rough correspondence between all the methods.

Acknowledgments

Conflict of Interests: None declared.

References

Hein, A.-M. K., Richardson, S., Causton, H. C., Ambler, G. K. and Green, P. J. (2005). BGX: a fully Bayesian gene expression index for affymetrix geneChip data. *Biostatistics* **6**, 349–73.

Hubbell, E., Lui, W. M. and Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics* **18**, 1585–92.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–64.

Li, C. and Wong, W. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences* **98**, 31–6.

Liu, X., Milo, M., Lawrence, N. D. and Rattray, M. (2005). A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics* **21**, 3637–44.

Milo, M., Fazeli, A., Niranjan, M. and Lawrence, N. D. (2003). A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Society Transactions* **31**, 1510–2.

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* **99**, 909–17.

[*Received April 13, 2006; revised July 6, 2006; accepted for publication August 2, 2006*]